Harmonic Detection[*]

Robert Ahlfinger
Brenton Cheeseman
Patrick Doody

Abstract

This is a discussion as to the ideas and methods of finding the harmonics in a speech signal.

1 The Biggest Obstacle

There is no question about it. For this algorithm to work correctly, the obstacle that is simultaneously most critical and most prone to error is accurately and consistently detecting the first harmonic in a chunk of speech. For instance, if the software incorrectly thinks the person speaks with a very deep voice in a particular chunk, the resulting frequency shift to the actual first harmonic will be enormous. The ratio of the correct index to the approximated index of the first harmonic is equal to the ratio of the actual shift in pitch and the desired shift in pitch after the voice manipulation is complete.

2 A Brief Overview of Harmonics and Speech

Why does middle C sound different from a piano, a trumpet, or an opera singer? After all, they all have the same pitch. The difference rests not in the base frequency that is being played per se, but rather in the sound's harmonics. Whenever an instrument (or a voice) makes a sound, the pitch you hear is called the first harmonic, it is the lowest and usually the strongest frequency emitted. However, this is not the only noise that is produced. There are also waves produced at all the higher octaves on the same note. The sound produced exactly one octave higher than the first harmonic is the second harmonic, the next octave up is the third harmonic, and so on. Looking at the Fourier Domain, it is important to remember that each octave, and therefore each harmonic, is exactly twice the frequency of the one below it. The relative strength or weakness of each individual harmonic gives each instrument a unique sound. In the case of speech, our vocal cords determine the pitch and produce the harmonics while our mouths individually dampen each harmonic in a set pattern to make a particular vowel. Consonants, unlike vowels, do not have a pitch nor do they have harmonics. A person's articulation of an 's' or 'z' sound, for instance, does not change depending on whether or not he has just been kicked in the groin.

[*] Version 1.5: Aug 12, 2005 4:20 pm GMT-5

[†] http://creativecommons.org/licenses/by/1.0

Spectrum For A Chunk

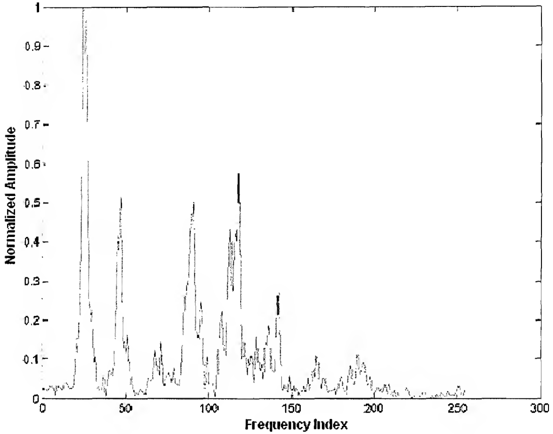


Figure 1: DFT of one 512 sample chunck of a speech signal.

3 Multitasking

Because consonants (along with periods of silence or noise) do not have pitch, our harmonic detection algorithm has the double duty of determining if a vowel noise is being produced in the first place, and if so, the location of the first harmonic as well. If a 'k' sound is mistaken for a vowel, for instance, the pitch synthesizer would attempt to shift its frequencies up the spectrum, resulting in a nasty high frequency noise that would not be mistaken for a 'k'.

4 A Naïve Approach

Before hitting gold, we developed several techniques to do this job that all fell short of satisfaction. One such technique was to construct a zero padded vector equal to the length of the DFT that had ones only at multiples of an integer that was a candidate for being the location of the first harmonic. After taking a dot product of these two vectors, we would try again for a different candidate index. The thought was that the largest resulting dot product would correspond to the correct placement of harmonics since they lined up with the largest values in the spectrum. However, if the harmonics do not appear at exact multiples of the candidate integer, this technique is worthless. Too much noise ruins its effectiveness as well.

DFT and an Example Comparison Vector

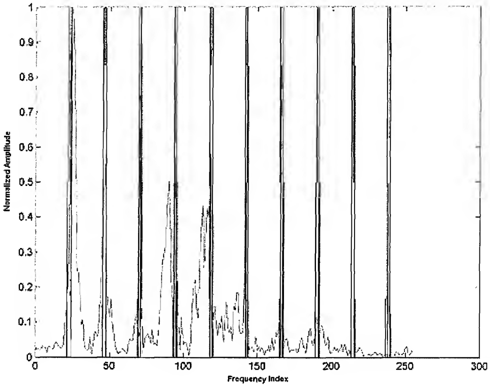


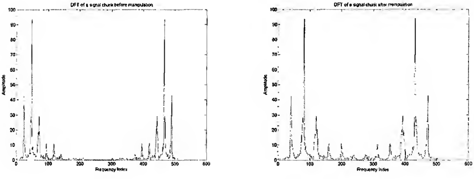
Figure 2: DFT of one 512 sample chunck of a speech signal overlapped with a comparison vector of about 24Hz, illustrating the general location of the harmonics.

To get rid of the first problem, we started using vectors that had a window of three ones around integer multiples of the candidate to allow some wiggle room for the actual location of the higher harmonics. Finally, we tried taking the logarithm of the values in the spectrum with the hope that the borders of the harmonics would stick up much farther than adjacent frequencies. If this held true to a greater extent than any other random locations in the spectrum, we could isolate the harmonics with the right type of high pass filter. In the end, we discovered each of these techniques were pretty good at finding harmonics in a certain kind of spectrum and failed miserably in other conditions. We needed something that worked all the time.

5 Hitting the Jackpot

The algorithm that works far and away better than any others we tested relies on the principle that the DFT of a chunk, like the time domain version of the chunk itself, has non-periodic and periodic aspects. In the first half of the DFT, the only repetition comes from the evenly spaced peaks of the harmonics. Everything else, whether noise or spectrum elements resulting from a consonant, is not periodic. Therefore, we take the first half of the magnitude of our DFT as a new signal to look at. Naturally, to analyze it we take the DFT of this vector, and look at the magnitude of the result. So now we have the tongue twisting magnitude of the DFT of the magnitude of the first half of the DFT of the original signal chunk. The DFT of the DFT!

DFT of Signal Sample



(a) (b)

Figure 3: (a) Discrete Fourier Transform showing the spectra for one 512 sample chunk of the speech signal before manipulation by the Pitch Synthesizer. (b) Discrete Fourier Transform of the original DFT spectra for one 512 sample chunk of the speech signal after manipulation by the Pitch Synthesizer.

The new spectrum invariably contains a very large DC value and a lot of power on the low end of the spectrum resulting from the necessarily positive average value of a magnitude plot (remember we used the magnitude of the original DFT) along with non-periodic elements from noise or consonants. But for n greater than two or three, this new DFT goes straight to zero and stays there until it hits the only periodic element of the original DFT –the harmonics. By ignoring the first couple of values on our new spectrum, we very accurately find the first harmonic by taking the first frequency with a magnitude that is on par with the large DC value. If no such frequencies exist, we can safely assume that the chunk does not contain a vowel and does not need manipulation. This new sneaky trick (taking the DFT of the DFT) is very precise and extremely consistent, especially in the presence of noise. In fact, had we discovered this earlier, there is probably another whole project in developing this particular tool in much greater depth. It could be used to automatically detect different types of human sounds, such as separate voiced and unvoiced fricative sounds as well as the tried and true vowels. Another use would be to compute the signal to noise ratio without having access to the original signal and figuring out whether the signal chunk should even be considered worthy of processing because of the prevalence of noise.